Module 2: Yelp Sentiment Analysis

Group 4

Sam W., Hongqian X., Ke T.

University of Wisconsin-Madison

March 7th, 2019

Overview

1. Plan
   - Data Cleaning
   - Natural Language Processing
   - Modeling & Analysis

2. Preliminary Analysis
   - Data Visualization
   - Phrases extraction

Plan

Plan: Data Cleaning

- JSON objects $\rightarrow$ Tabular CSV
  - For ~ 6.8 million JSON objects, this takes about 20 minutes
  - Also need to "flatten" some columns
- Consolidate categories
  - Dataset contains over $1,000$ unique categories
  - Keep only the top 30 categories, label the rest as **Other**
- Missing values not a major source of concern

Plan: Natural Language Processing

- Purpose: convert a text to a vector
- Process:
  - Split sentences into lists of words and noun phrases
  - Calculate tf-idf value for each word and noun phrases
  - Choose a bunch of words and noun phrases as features, take tf-idf value as feature value

Plan: Natural Language Processing

Split sentences into lists of words and noun phrases

Original	Removing numeric and punctuation characters	Lowercase and split	Remove stop words	Reduce words to their root forms
Total bill for this horrible service?	Total bill for this horrible service	['total', 'bill', 'for', 'this', 'horrible', 'service']	['total', 'bill', 'horrible', 'service']	['total', 'bill', 'horribl', 'servic','horrible service']

Plan: Natural Language Processing

- Calculate Term Frequency-Inverse Document Frequency :

$$tf\text{-}idf(t, d, D) = tf(t, d) \times idf(t, D)$$

$$tf(t, d) = \frac{Number\ of\ times\ term\ t\ appears\ in\ a\ document\ d}{Number\ of\ all\ terms\ in\ document\ d}$$

$$idf(t, D) = \log \frac{Number\ of\ all\ documents\ D}{Number\ of\ documents\ with\ term\ t\ in\ it}$$

Plan: Natural Language Processing

Example:
review$_1$: "The hotel is horrible!"
review$_2$: "What a great hotel!"

tf-idf for terms in review$_1$			
term	tf	idf	tf-idf
hotel	1/4	0	0
horrible	1/4	log2	$1/4 \times log2$

- Both words have the same term frequency
- **hotel** is penalized for appearing in both reviews
- **horrible** has a higher score because it only appears in one of the reviews

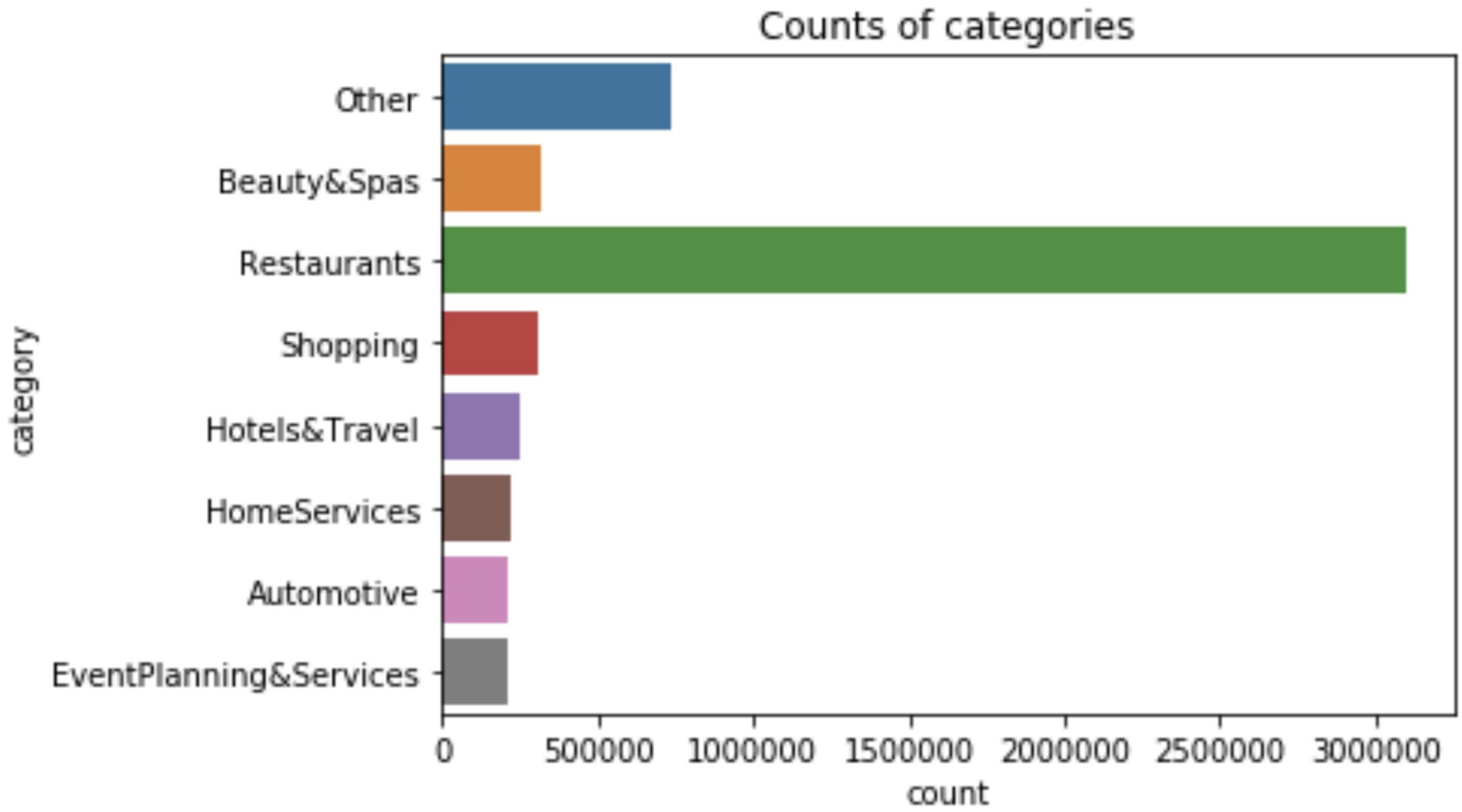
Plan: Modeling & Analysis

- Plan to try multiple models and compare results
- Model requirements:
  - Interpretable
  - Capable of handling high-dimensional dataset
  - Relatively accurate
- Model works $\rightarrow$ use feature importance to make recommendations
  - Otherwise, revise our NLP approach
- Create charts/visual evidence to support our findings

Preliminary Analysis

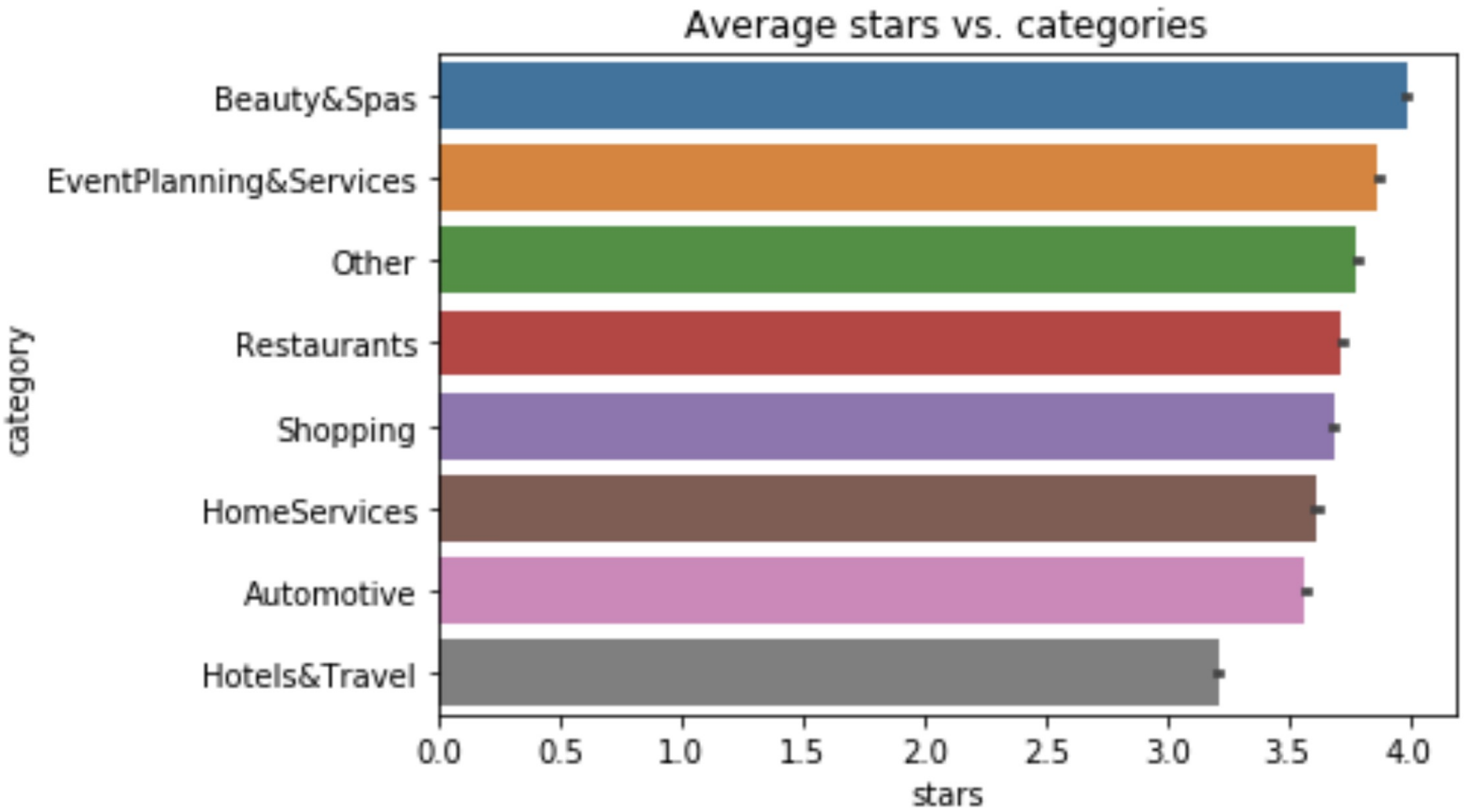
Counts of Categories

- **Restaurant** is the one with the most reviews.



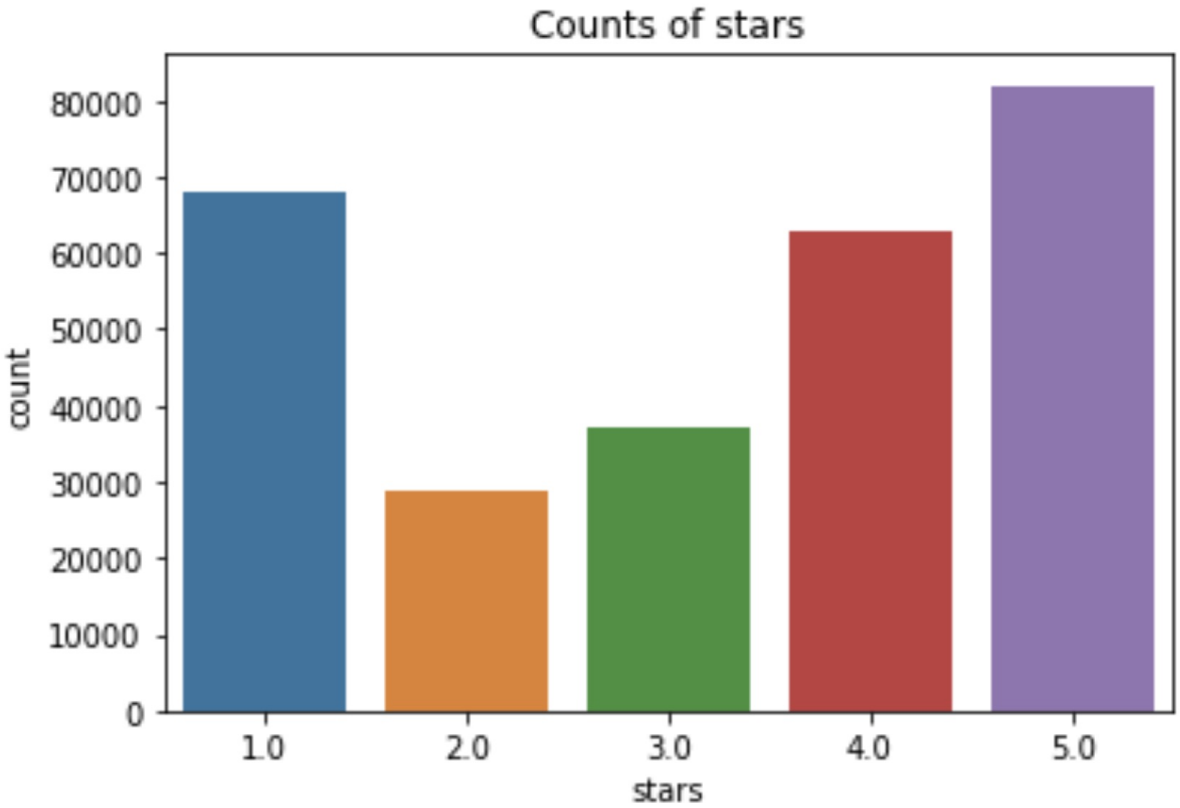
Average Stars vs. Categories

- **Hotel&Travel** is the one with the lowest average stars.



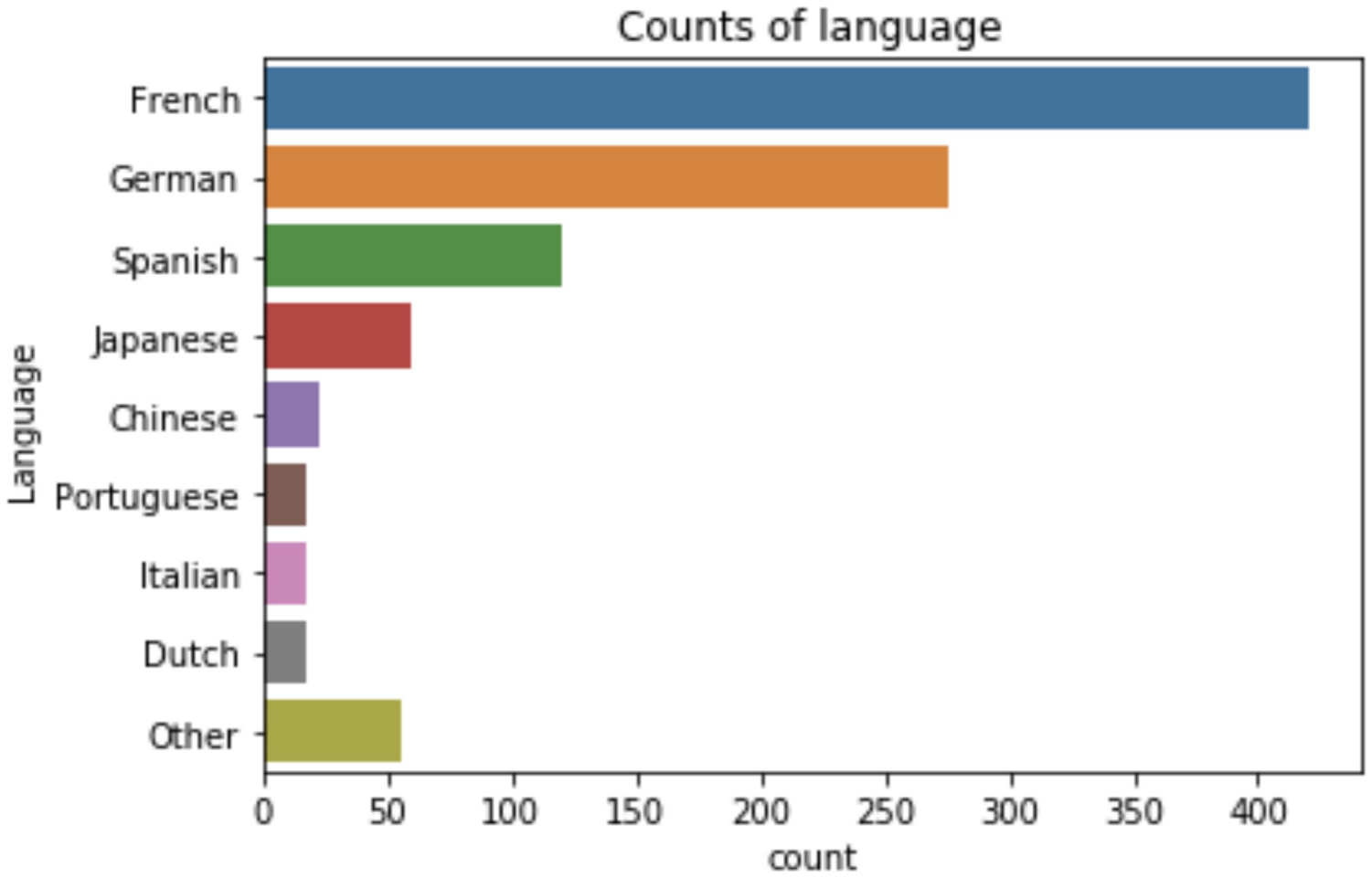
Hotel: Counts of Stars

- Number of hotels: 4833;
- Number of reviews: 278733.



Hotel: Counts of Language

- 23 kinds of foreign languages;
- Number of reviews in foreign language: 1006.



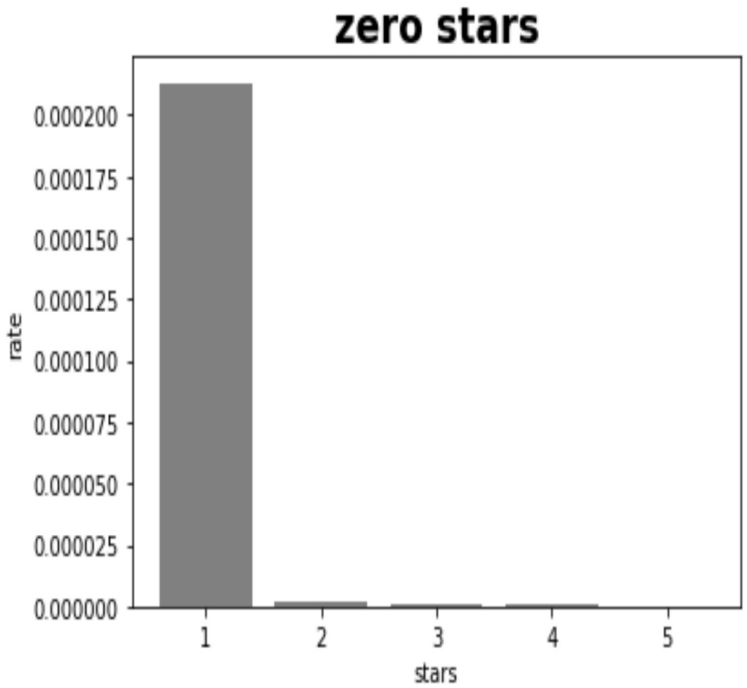
Hotel: Top 10 Negative Noun Phrases

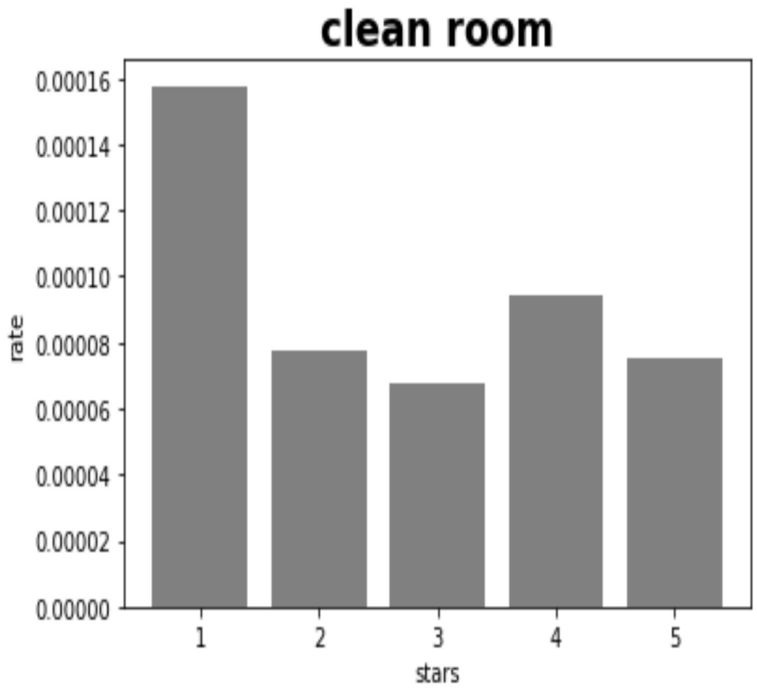
	1	2	3	4	5
front desk	4028	1743	1149	1058	791
customer service	2111	395	257	281	646
credit card	885	191	110	86	80
las vegas	831	303	387	844	1474
resort fee	699	526	500	529	252
zero stars	696	6	4	2	0
new room	606	211	113	89	52
rental car	545	119	109	142	254
clean room	515	253	222	307	246
room service	442	307	326	504	482

Figure: Counts of noun phrases

Hotel: Negative Noun Phrases Comparison

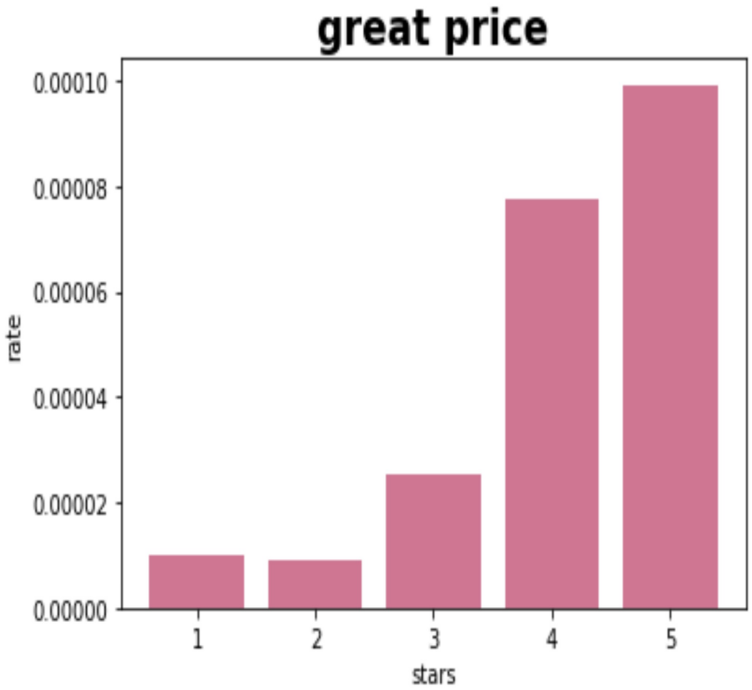
- **zero stars** is predictive but not useful for making suggestions
- **clean room** is less predictive but useful for making recommendations

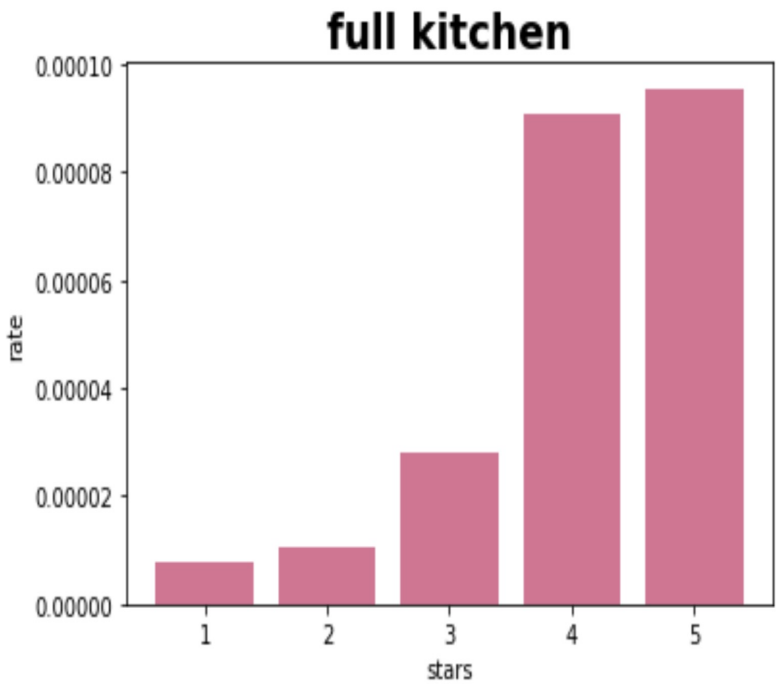




Hotel: Positive Noun Phrases Comparison

- **great price** means that customers like a low price
- **full kitchen** suggests that customers appreciate food in their room





The End